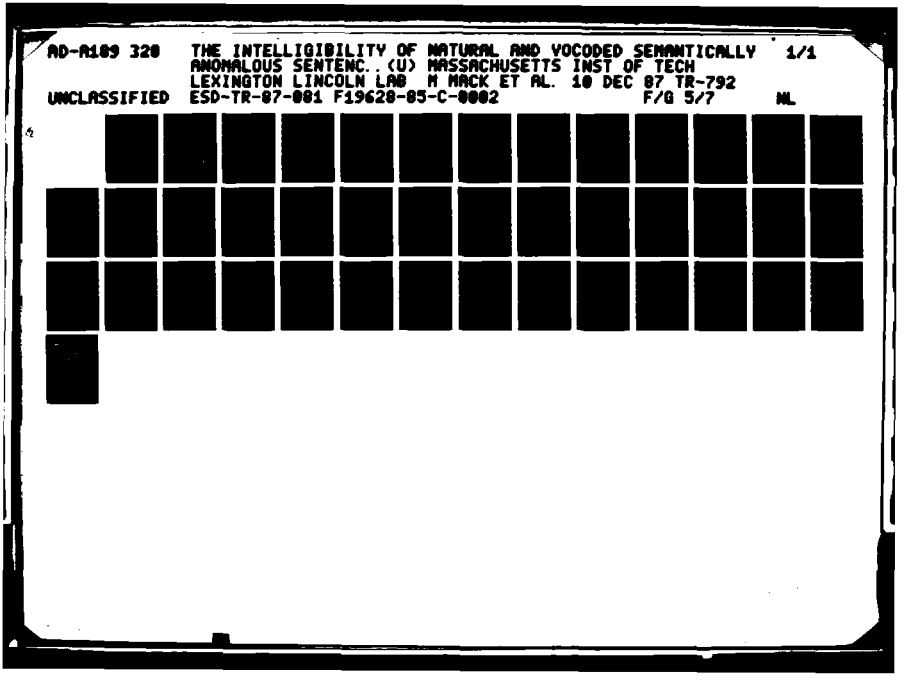
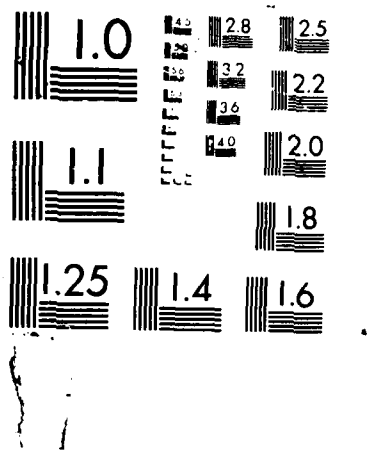
AD-A189 320 THE INTELLIGIBILITY OF NATURAL AND VOCODED SEMANTICALLY ANOMALOUS SENTENC..(U) MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB M MACK ET AL. 10 DEC 87 TR-792 1/1

UNCLASSIFIED ESD-TR-87-081 F19628-85-C-0002 F/G 5/7 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

ESD-TR-87-081

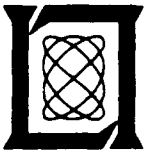
Technical Report
792

The Intelligibility of Natural and Vocoded Semantically Anomalous Sentences: A Comparative Analysis of English Monolinguals and German-English Bilinguals

M. Mack
J. Tierney

10 December 1987

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

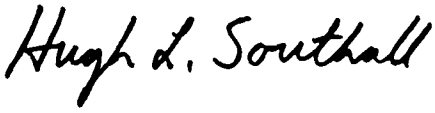


Prepared for the Department of the Air Force
under Electronic Systems Division Contract F19628-85-C-0002.

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the Department of the Air Force under Contract F19628-85-C-0002.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER



Hugh L. Southall, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

THE INTELLIGIBILITY OF NATURAL AND VOCODED SEMANTICALLY ANOMALOUS SENTENCES: A COMPARATIVE ANALYSIS OF ENGLISH MONOLINGUALS AND GERMAN-ENGLISH BILINGUALS

M. MACK
University of Illinois

J. TIERNEY
Group 24

TECHNICAL REPORT 792

10 DECEMBER 1987

LEXINGTON MASSACHUSETTS

ABSTRACT

The present study was undertaken in order to analyze the performance of 24 German-dominant German-English bilinguals and 24 English monolinguals on tests of semantically anomalous natural and computer-generated (vocoded) sentences. The primary objectives of the study were these: to determine whether the overall performance of the bilinguals was significantly worse than the monolinguals' in response to natural and/or vocoded speech; to categorize the specific types of errors made by the two groups of subjects; and to analyze the patterns of errors made by the two groups in an attempt to evaluate their sentence-processing strategies. Secondary objectives of the study were these: to assess the relationship between the groups' subjective ratings of task difficulty and their test performance; and to assess the relationship between the bilinguals' subjective evaluation of their English proficiency and their test performance.

Results revealed that, overall, the bilinguals made more errors than the monolinguals did in response to both natural and vocoded speech. Further, both groups had a preponderance of phonemic rather than morpho-syntactic or lexico-semantic errors, and most errors were phonemic substitutions, rather than omissions or insertions. Results of the analysis suggested that the bilinguals and monolinguals were using similar processing strategies. However, the bilinguals' pattern of errors indicated that they found both the natural and vocoded sentences tasks considerably more difficult than the monolinguals did, and the number of overall errors in the vocoded sentence task indicated a potential problem for communication systems of this type. It was also found that subjective task-difficulty ratings and English-proficiency ratings were correlated with test scores.

Practical and theoretical implications of the results are considered, and suggestions for additional research are provided.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

THE INTELLIGIBILITY OF NATURAL AND VOCODED SEMANTICALLY ANOMALOUS SENTENCES: A COMPARATIVE ANALYSIS OF ENGLISH MONOLINGUALS AND GERMAN-ENGLISH BILINGUALS

1. INTRODUCTION

It is widely accepted that natural speech is more intelligible than synthetic speech, even when synthetic speech is of relatively high quality. The superiority of natural over synthetic speech has been demonstrated by a number of researchers utilizing various experimental instruments, including the Diagnostic Rhyme Test (DRT) and the Modified Rhyme Test (MRT), as well as tests of lexical decision, word recall, sentence recall, and sentence verification [1, 2, 3, 4, 5, 6, 7].

Needless to say, subjects' performance in response to synthetic versus natural speech is highly dependent upon the demands of the perceptual task, the characteristics of the stimuli, and the quality of the speech synthesizer. Hence, it is difficult to specify, in absolute terms, the magnitude of the decrement in perceptual performance when subjects are presented with synthetic versus natural speech under different test conditions. Yet a consistent and general finding of studies in this area has been that subjects' performance in response to natural speech is superior to their performance in response to synthetic speech although, at least in some instances, this superiority has apparently been slight [5, 8]. Pisoni *et al.* [7] have suggested that some behavioral measures used to assess intelligibility have simply been "too gross and insensitive to reveal differences between various types of speech, p. 22."

Hence, it is appropriate to explore speech perception tasks which place moderate to high demands on subjects' processing abilities. It is also appropriate to examine various scoring procedures. For example, a study by Mack and Gold [9] utilized semantically anomalous sentences in a sentence-transcription task. The Mack and Gold study revealed that natural and vocoded sentences yielded nearly identical intelligibility scores when these scores were based on the percentage of words rendered correctly [9]. However, a larger difference emerged when *all* errors were tabulated (thereby including multiple errors within a word). Using the latter metric, it was found that vocoded sentences resulted in 2.5 times as many errors as natural sentences did. In addition, it was found that, in the vocoded condition, over 40% of all errors were *not* phonemic, but were morpho-syntactic or lexico-semantic. This is of special interest in light of the fact that many intelligibility tests are designed to evaluate perceptual accuracy only in terms of segmental phonemic properties.

Of equal interest is the fact that nearly all speech intelligibility results obtained to date (at least in the United States) apply only to native speakers of English, since very few synthetic-speech experiments have used non-native speakers as subjects. There are, however, compelling reasons for undertaking studies of such subjects: Previous experimental work has revealed that

even fluent bilinguals may exhibit perceptual patterns which are unlike those of monolinguals [10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. There is also evidence that bilinguals who are not native speakers find it more difficult to process distorted or degraded speech than do monolinguals [20, 21].

These findings have serious ramifications for the field of speech synthesis and communication. Indeed, there are numerous occasions in which non-native speakers of English are required to receive English messages over noisy or degraded channels. An example would be the case in which a German pilot must receive vocoded messages in English. Although the pilot may be highly fluent in English, the fact that he is a non-native speaker receiving vocoded speech in ambient noise could have grave consequences if he were to experience even moderate difficulty in understanding the speech conveyed to him. So, from a practical standpoint, it is evident that careful research into the perceptual performance of non-native speakers confronted with synthetic speech is badly needed.

But aside from its obvious practical import, such research also has major theoretical implications. For example, detailed comparative analyses of the performance of native and non-native speakers makes it possible to assess the limits on speech intelligibility imposed by synthetic speech; to determine precisely how non-native speakers' performance in response to synthetic speech differs from the performance of native speakers; and to gain insight regarding the strategy (e.g., "top-down" versus "bottom-up") used by non-native speakers (or, more properly, non-native listeners) when they attempt to process synthetic speech.

Therefore, in the present experiment, the object of study was the performance of native and non-native speakers of English who were required to transcribe a set of 57 semantically anomalous natural and synthetic (vocoded) English sentences. In order to gain insight into the precise nature of the subjects' errors and processing strategies, highly detailed error analyses were utilized. The major questions addressed were these: (1) Is the overall perceptual performance of non-native speakers of English significantly worse than that of native speakers in response to natural and/or vocoded semantically anomalous sentences? (2) What specific types of error patterns may be observed in the responses of native and non-native speakers to such sentences? (3) Do these error patterns reflect differences between native and non-native speakers' processing strategies?

A secondary objective of the study involved examining the relationship between subjects' test scores and selected subjective measures in order to determine the validity of subjective evaluations in experiments of this type. Specifically, the questions addressed were these: (1) Is there a correlation between subjects' task-difficulty assessments and their test scores? (2) Is there a correlation between bilinguals' self-evaluated English proficiency and their test scores?

2. EXPERIMENT

2.1 SUBJECTS

Forty-eight subjects participated in the present experiment. This included 24 English monolinguals and 24 German-English bilinguals. Within each of these groups of 24, 12 subjects were randomly assigned to one of two listening conditions, natural speech and vocoded speech, so that there were 4 groups of 12 subjects. Subjects ranged in age from 20 to 53. Of the monolinguals, 13 were female and 11 were male; of the bilinguals, 10 were female and 14 were male. All subjects were students, staff, or faculty members at the University of Illinois at Urbana. Subjects' participation was voluntary, and all subjects were paid $5.

Demographic and language-acquisition information about the German-English bilinguals was obtained via a questionnaire nearly identical to the one used in [22] (Appendix I, Part 1). Responses to this questionnaire revealed that 23 of the 24 bilinguals were citizens of either West Germany or Austria, while one was a U.S. citizen. The native language (L1) of all was German, and all were raised in German-language homes with parents whose native language was German. Seventeen of the subjects reported that they spoke at least one language other than English. The preferred language of 14 of the bilinguals was German while for 3 it was English and for 1 it was French. Six stated that their language preference was dependent upon linguistic context. (It should be noted that language preference and language dominance are not necessarily equivalent.) Twelve of the subjects indicated that they counted in German, while 16 indicated that they said the alphabet in German, and 10 indicated that they spoke to themselves in German. These three linguistic tasks — counting, saying the alphabet, and using internal speech — are usually interpreted as automatic speech processes in the L1, and thus are especially resistant to change due to the influence of the second language (L2). In addition, it was found that the bilinguals' mean age at the onset of their acquisition of English was 10 years and 9 months with a range of 5 to 23. Twenty-one of the 24 bilinguals had begun their acquisition of English by the age of 13. All subjects had completed high school and most were currently enrolled in college as undergraduate or graduate students.

Bilingual subjects also completed a self-evaluation language-proficiency questionnaire (Appendix I, Part 2). The items used in this questionnaire represented a subset of those used in a questionnaire originally developed by the Foreign Service Institute to help assess L2 proficiency. A scoring procedure identical to that used by Mack [22] to assess self-evaluation among early French-English bilinguals was employed in order to obtain a quantitative measure of the subjects' perceived fluency in English.* The highest score achievable was 23, the score a native speaker of

* The subjective self-evaluation portion of the questionnaire was scored as follows: Each "no" response to questions 1-6 and to question 10 received 1 point, and each "yes" response to questions 7-9 and question 12 received 1 point. To question 11, the responses "never", "sometimes", and "always" were given point values of 0, 1, and 2, respectively. For question 13, the point value was that provided by the subject. Thus, the higher the score, the greater was the subject's self-evaluated fluency in English. Because all subjects were native speakers of German with professed complete fluency in the language, they were not given language-background questionnaires for German.

English would be expected to obtain. The bilinguals in the present study obtained a mean fluency rating of 14.48. The fluency ratings were later used in a correlational analysis designed to determine the strength of the relationship between the bilinguals' self-evaluated English proficiency and their actual English scores. (To determine that the bilingual subjects in the natural and vocoded groups were essentially equivalent with respect to their English-language ability, a t-test of independent means was carried out on their self-evaluation scores. It revealed no significant difference in the average fluency scores of the bilinguals assigned to the two groups.)

Finally, a pretest of meaningful natural English sentences was administered to all subjects to determine that they could perform adequately in the experiment and to determine whether or not the bilinguals in the natural-speech group and in the vocoded-speech group were equivalent. Analysis of this pretest resulted in the rejection of 3 bilinguals whose error rates were greater than 3 standard deviations below the mean of their test group. These 3 were later replaced by bilinguals whose performance reached criterion on the pretest. Evidence from the pretest also indicated that the bilinguals assigned to the two test conditions were equivalent in their ability to perceive and accurately render English sentences: The bilinguals assigned to the natural-speech condition had an average of 95.0% words correct, while the bilinguals assigned to the vocoded-speech condition had an average of 95.6% words correct.

2.2 STIMULI

Stimuli consisted of 57 semantically anomalous sentences identical to those used by Mack and Gold [9]. These sentences appear in Appendix II. Sentential rather than isolated-word stimuli were used so that subjects would be required to utilize some of the same processing mechanisms used in ordinary communication. For example, it is well known that words produced fluently in sentences are not identical to words produced in isolation [23, 24, 25, 26]. Further, sentences were anomalous rather than meaningful because it was believed that anomalous sentences would be more difficult to process and, hence, would be especially effective in revealing potential differences between the members of the language groups and test conditions.

Sentences were constructed with a set of relatively common nouns, adjectives, and verbs pseudo-randomly ordered to produce strings which were grammatical but meaningless. Word-initial consonants were phonemically balanced, with each one of 19 phonemes occurring 15 times — 6 times in nouns, 6 times in adjectives, and 3 times in verbs. The phonemes were /p, b, t, d, k, g, č, ǰ, š, f, v, s, z, m, n, l, r, θ, h/. All sentences were of the form S → NP + VP where NP → (det.) + adj. + noun and VP → verb + det. + adj. + noun. Sixteen of the sentences had no determiner in sentence-initial position. The same sentences were used in both the natural and vocoded conditions. Each set of 57 sentences contained 383 words. Hence, 18,384 words (383 words × 48 subjects) were analyzed in detail.

The anomalous sentences were tape recorded by the experimenter (MM) at a normal speaking rate. The sentence onsets were separated by 20-second intervals. For the vocoded condition, the tape recording of the sentences was used as input to a real-time channel vocoder with an 8000 bps data rate [5], used in earlier work [9]. At this data rate, channel vocoder transmission is almost of telephone quality. The Lincoln Digital Signal Processors (LDSPs), simple programmable computers of a Harvard architecture, were used to implement the vocoder program. Sentences were sampled at 10 kHz and then, as they were vocoded, they were recorded on reel-to-reel audio tape for later presentation.

2.3 PROCEDURE

Subjects were tested in groups of varying sizes in the language laboratory at the University of Illinois. Sentences were presented on a Tandberg TB5200 tape recorder whose output was directed to individual listening consoles. Subjects heard the sentences over Tandberg stereo headphones with amplitude set at a comfortable listening level. Subjects could alter the amplitude at their consoles if they wished.

Subjects were given answer booklets in which to write the test sentences. They were told to write each sentence as accurately as possible, as soon as it was presented, and to guess if they were uncertain. Following completion of the test, subjects gave it a difficulty rating on a scale of 1 (extremely easy) to 10 (extremely difficult). The entire test session lasted about 45 minutes.

2.4 DATA ANALYSIS

Orthographic transcriptions of all the words were made by the experimenter (MM). The data were analyzed in various ways, including a tabulation of *all* errors (which could include more than one error per word) and of all incorrectly rendered words. The latter tabulation was the basis for determining total percent correct. In addition, each error was analyzed as being "linguistic," "transpositional," or "other." Analysis was also made to determine the position, within each stimulus sentence, of erroneous words. In the error analyses, misspellings were not counted as errors, provided that they seemed to reflect an attempt to render the sounds as they were heard. Thus, spellings such as <reddy> for "ready" and <tailer> for "tailor" were considered entirely acceptable, as was <latter> for "ladder."

Procedures for conducting these analyses are discussed in greater detail in sections 2.4.1 through 2.4.6 below.

2.4.1 Overall Errors

For this analysis, *every* error was counted. Thus, it was possible for a single word to exhibit more than one error. For example, if the word "Tim" was rendered as "kin," two errors were counted, one in word-initial and one in word-final position.

2.4.2 Linguistic Errors

A linguistic error was phonemic, morpho-syntactic, or lexico-semantic. A phonemic error was classified as one in which a single phoneme or cluster was involved (e.g., "thimble" → "symbol"; "lean" → "green"). A morpho-syntactic error was classified as one involving a derivational or inflectional morpheme (e.g., "liked" → "likes") or a function word (e.g., "ϕ" → "to"). A lexico-semantic error was classified as one in which the target word and the word provided by the subject bore some relation in meaning. This relation was most often one of similarity or synonymy (e.g., "first" → "third"; "careful" → "cautious").

2.4.3 Transpositional Errors

A transpositional error was one of substitution, omission, or insertion. A substitution error involved the confusion of one phoneme, bound morpheme, function word, or content word for another (e.g., "thimble" → "symbol"; "liked" → "likes"; "first" → "third"). An omission error arose when any of the above was omitted (e.g., "master" → "aster"; "newer" → "new"; "chief" →"ϕ"). An insertion error occurred when any of the above was inserted (e.g., "raid" → "grade"; "jewel" → "jewels"; "ϕ" → "time").

As is apparent, all transpositional errors could be cross-classified with linguistic errors. For example, the rendering of "symbol" for "thimble" reflected both a substitution and a phonemic error.

2.4.4 "Other" Errors

Errors which did not fall into one of the three linguistic categories were classified as "other." This category included errors of perseveration (the repetition of a word which appeared in the preceding sentence or earlier in the same sentence) and errors of metathesis (e.g., "shines a safe" → "saves a shine"), as well as inexplicable errors (e.g., "Tim" → "camera").

2.4.5 Position-in-Sentence Errors

For the purposes of this analysis, all words which contained at least one error were counted and note was made of their position within the sentence. Thus, if a subject rendered the stimulus sentence, "A paper nature seeks the cool master," as "The papered nature seeps the cool aster," an error was counted for positions 1, 2, 4, and 7. For six-word sentences, position 1 was considered null.

Although this analysis is termed a "position-in-sentence" analysis, it could not be determined with certainty whether it was position or part of speech which was most directly responsible for the pattern of errors, since each within-sentence position was also associated with a particular part of speech.

2.4.6 Statistical Analyses

For the analysis of raw error data, a 4-way repeated-measures analysis of variance with two between and two within factors was used [27]. This ANOVA had a $2 \times 2 \times 3 \times 3$ design (language group $\times$ speech type $\times$ linguistic-error type $\times$ transpositional-error type). In addition, a repeated-measures analysis of variance was conducted on the "other" errors. A Chi square (χ^2) procedure was conducted in the analysis of the total number of words rendered correctly by the two language groups in the two speech conditions. For the analysis of proportional data in the examination of the relative distribution of error types, a z-test of the significance of the difference between two independent proportions was used [28]. Post-hoc tests were conducted using the Tukey HSD procedure with α set at 0.05 [29].

2.5 RESULTS

2.5.1 Overall Errors

The bilinguals made far more overall errors in response to both the natural and to the vocoded sentences than did the monolinguals, as Table I and Figure 1 reveal. In fact, the bilinguals made over 11 times as many errors as the monolinguals did in response to natural speech, and nearly 3 times as many in response to vocoded speech. For both groups, vocoded speech resulted in more errors than natural speech did. However, there was a relatively small difference between the number of errors made by the bilinguals in the natural and vocoded conditions, for they made, on the average, only 1.25 times as many errors on vocoded as on natural speech. The difference was much larger for the monolinguals who made, on the average, over 5 times as many errors on vocoded as on natural speech.*

Statistical analysis revealed one highly significant main effect for language group [$F(1, 44) = 62.78$, $p < 0.0001$], with the bilinguals' performance being worse than the monolinguals, and for speech type [$F(1, 44) = 7.85$, $p < 0.008$], with vocoded speech causing more errors than natural speech. There was no significant language-group by speech-type interaction.

Subjects' overall performance was also quantified in terms of percentage of correct words — a value based upon the total number of words rendered correctly divided by the total number of words in the stimulus set. The monolinguals in the natural and vocoded conditions had an average of 98.72% and 92.01% words correct, respectively, while the bilinguals in the natural and vocoded conditions had 85.29% and 81.38% words correct, respectively (see Figure 2).

---

* In Mack and Gold [9], English monolinguals made 2.5 times as many errors on vocoded as on natural sentences. Because the sentence stimuli in that experiment were identical to those used in the present one, it was not readily apparent why the monolingual subjects in the present experiment had *5 times* as many errors on vocoded as on natural sentences. One possibility is that, because the subjects in [9] were employees of the MIT Lincoln Laboratory, some may have been familiar with vocoded speech. This familiarity may have rendered vocoded speech more intelligible to them than it was to the present group of "naive" listeners.

TABLE I Overall Errors				
	Monolinguals		Bilinguals	
	Natural	Vocoded	Natural	Vocoded
Total	71	367	799	998
Mean	5.92	30.58	66.58	83.17
S.D.	4.10	14.34	34.44	33.29
Range	0-12	11-65	3-122	41-129

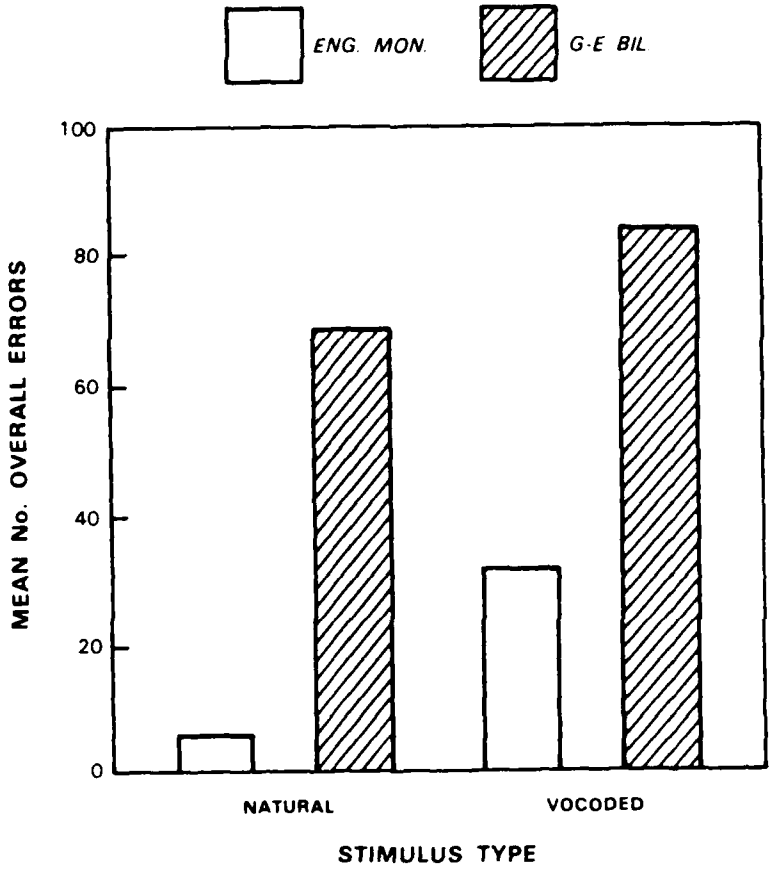


Figure 1. Comparison of overall errors.

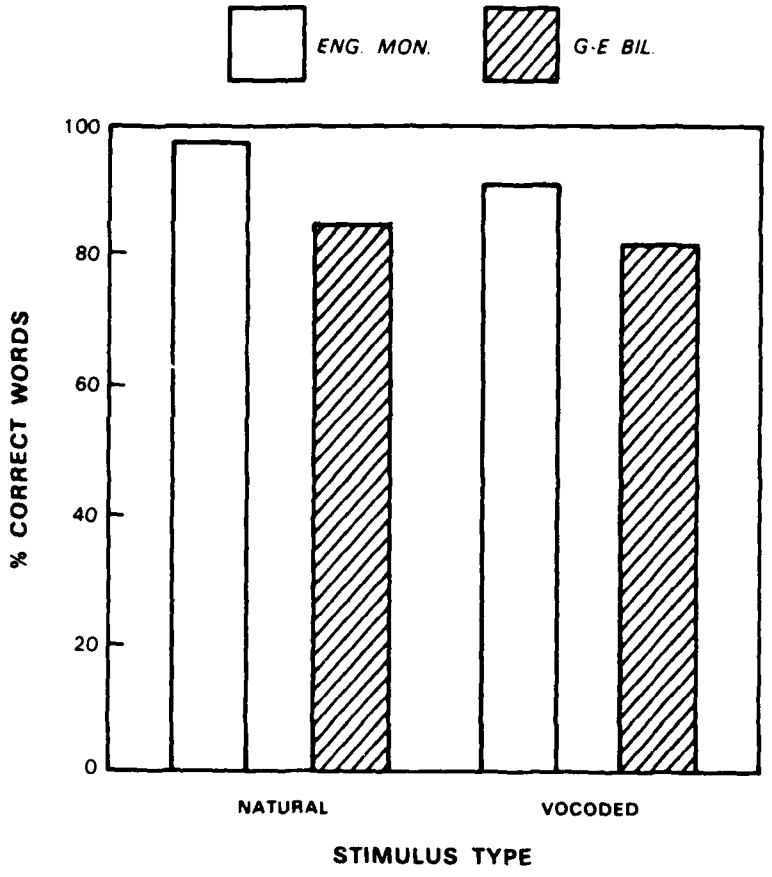


Figure 2. Comparison of percentage of correct words.

A Chi square (χ^2) test revealed a highly significant difference between the overall number of words rendered correctly by the monolinguals and the bilinguals in the natural condition ($\chi^2 = 564.08$, $p < 0.0001$). There was likewise a highly significant difference between the overall number of correct words for the monolinguals and the bilinguals in the vocoded condition ($\chi^2 = 225.52$, $p < 0.0001$).

Thus, in terms of overall errors and percentage of correctly rendered words, the bilinguals performed significantly worse than the monolinguals in response to natural and vocoded speech.

2.5.2 Linguistic Errors

For both groups of subjects, the largest number of linguistic errors were phonemic; approximately 60-70% of all linguistic errors fell into this category (see Table II). Nonetheless, for all groups of subjects, a substantial percentage of errors (about 30-40%) were not phonemic, but were morpho-syntactic and lexico-semantic.

Statistical analysis revealed a highly significant main effect for linguistic errors [$F(2, 88) = 67.75$, $p < 0.0001$]. Post-hoc analysis indicated that, overall, there were significantly more phonemic errors than morpho-syntactic or lexico-semantic errors [HSD (0.05) = 1.57].

There was also a significant language group by linguistic-error interaction [$F(2,88) = 26.31$, $p < 0.0001$]. Post-hoc analysis revealed that there were significantly more phonemic than lexico-semantic errors [HSD (0.05) = 2.73] for both groups of subjects. The bilinguals (but not the monolinguals) also had significantly more phonemic than morpho-syntactic errors [HSD (0.05) = 2.73].

There was no significant speech type (natural versus vocoded) by linguistic-error interaction. Thus, vocoded speech did not have a significant effect on any *specific* linguistic error type to a greater or lesser extent than natural speech did. Thus, the general pattern was that phonemic errors were most common and lexico-semantic errors least common for both natural and vocoded speech.

Table II also shows the relative distribution of linguistic error types for subject groups and speech types. (The relative distribution is represented as percentages.) It is apparent that the percentage of phonemic errors was largest for monolinguals and bilinguals in the natural condition.

TABLE II

Linguistic Errors:
Means and Percentages

	Monolinguals		Bilinguals	
	Natural	Vocoded	Natural	Vocoded
Phonemic	3.42 (68.40%)	17.67 (63.29%)	43.50 (70.54%)	45.08 (59.58%)
Morpho-Syntactic	1.33 (26.60%)	7.33 (26.25%)	9.17 (14.87%)	14.08 (18.61%)
Lexico-Semantic	0.25 (5.00%)	2.92 (10.46%)	9.00 (14.59%)	16.50 (21.81%)

Analysis of the significance of the difference between proportions revealed that the bilinguals had a significantly *smaller* proportion of morpho-syntactic errors than the monolinguals did in both the natural condition ($z = 2.41$, $p < 0.05$) and in the vocoded condition ($z = 2.60$, $p < 0.01$). They also had a significantly *larger* proportion of lexico-semantic errors than the monolinguals did in both the natural condition ($z = -2.06$, $p < 0.05$) and in the vocoded condition ($z = -4.54$, $p < 0.01$).

2.5.3 Transpositional Errors

For both groups of subjects, the largest number of transpositional errors were substitutions, with approximately 69-83% of the transpositional errors falling into this category (see Table III).

TABLE III

Transpositional Errors: Means and Percentages

	Monolinguals		Bilinguals	
	Natural	**Vocoded**	**Natural**	**Vocoded**
Substitution	4.17 (83.40%)	22.08 (79.11%)	45.25 (73.37%)	52.17 (68.94%)
Omission	0.33 (6.60%)	4.33 (15.51%)	13.42 (21.76%)	20.00 (26.43%)
Insertion	0.50 (10.00%)	1.50 (5.37%)	3.00 (4.86%)	3.50 (4.62%)

There was a highly significant main effect for transpositional errors [$F(2, 88) = 95.02$, $p < 0.0001$]. Post-hoc analysis revealed that, overall, there were significantly more substitution than omission or insertion errors [HSD (0.05) = 1.73]. There were also significantly more omission than insertion errors [HSD (0.05) = 1.73].

There was a significant language group by transpositional-error interaction [$F(2, 88) = 30.25$, $p < 0.0001$]. Post-hoc analysis revealed that there were significantly more substitution than omission or insertion errors for both groups of subjects [HSD (0.05) = 2.99]. Further, the bilinguals (but not the monolinguals) made significantly more omission than insertion errors [HSD (0.05) = 2.99]. For all groups, the fewest errors were insertions.

There was also a significant speech type (natural versus vocoded) by transpositional-error interaction [$F(2, 88) = 3.67$, $p < 0.0347$]. Post-hoc analysis indicated that both natural and vocoded speech yielded significantly more substitution errors than other transpositional errors [HSD (0.05) = 2.99]. Further, the number of substitutions in the vocoded condition was significantly greater than in the natural condition [HSD (0.05) = 2.99].

As Table III also reveals, substitution errors constituted the largest percentage of errors for the monolinguals and the bilinguals in both conditions.

A test of the significance of the difference between proportions revealed that the bilinguals had a significantly *larger* proportion of omissions than the monolinguals in both the natural

condition ($z = 2.47$, $p < 0.05$) and the vocoded condition ($z = -4.03$, $p < 0.01$). In addition, in the vocoded condition, the bilinguals had a significantly *smaller* proportion of substitutions than did the monolinguals ($z = 3.53$, $p < 0.01$).

2.5.4 Linguistic and Transpositional Errors: Additional Analysis

Because all transpositional errors co-varied with linguistic errors, cross-tabulations are provided in Table IV. (Note that, due to the very low frequencies in some of the categories,

TABLE IV

Linguistic and Transpositional Errors Cross-Tabulated: Totals*

	Monolinguals		Bilinguals	
	Natural	**Vocoded**	**Natural**	**Vocoded**
Phonemic				
Substitution	37	200	487	495
Omission	2	6	19	22
Insertion	2	6	16	24
Subtotal:	41	212	522	541
Morpho-Syntactic				
Substitution	11	60	50	127
Omission	1	16	41	25
Insertion	4	12	19	17
Subtotal:	16	88	110	169
Lexico-Semantic				
Substitution	2	5	6	4
Omission	1	30	101	193
Insertion	0	0	1	1
Subtotal:	3	35	108	198
Total	60	335	740	908

* "Other" errors are not included.

totals rather than means are provided.) The following general observations can be made: For both the monolinguals and the bilinguals, the *largest* number of errors were phonemic substitutions (e.g., /mIt/ → /nIt/) while the *fewest* number of errors were lexico-semantic insertions (e.g., ϕ → "love"). Among the bilinguals, lexico-semantic omissions were also numerous (e.g., "bossy" → ϕ). The bilinguals in the vocoded condition also had a fairly large number of morpho-syntactic substitutions (e.g., "a" → "the").

Three interactions were significant: linguistic-error type × transposition-error type [$F(4, 176) = 108.02$, $p < 0.0001$]; language group × linguistic-error type × transpositional-error type [$F(4, 176) = 45.63$, $p < 0.0001$]; and condition × linguistic-error type × transpositional-error type [$F(4, 176) = 3.94$, $p < 0.029$.

Post-hoc analysis of these interactions confirmed the following: (1) Overall, there were significantly more phonemic substitutions than any other error type [HSD (0.05) = 3.57]; (2) the number of phonemic substitution errors made by the bilinguals was significantly greater than *any* other error type made by the bilinguals or monolinguals [HSD (0.05) = 5.14]; (3) the bilinguals had significantly more lexico-semantic omissions than the monolinguals did [HSD (0.05) = 5.14]; and (4) vocoded speech resulted in significantly more phonemic and morpho-syntactic substitution errors than natural speech did [HSD (0.05) = 5.14].

2.5.5 "Other" Errors

Errors which fell into the category "other" were not, it will be recalled, cross-classified as linguistic or transpositional errors. As may be seen in Table V, there were very few of these errors.

TABLE V

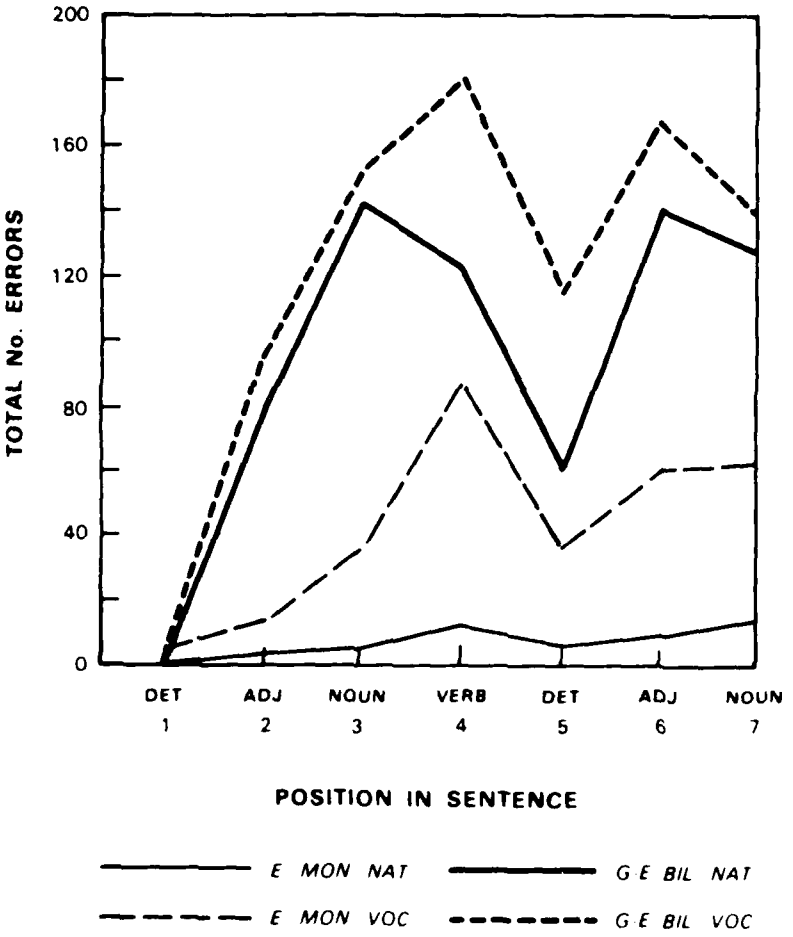
"Other" Errors: Means and Percentages

	Monolinguals		Bilinguals	
	Natural	Vocoded	Natural	Vocoded
Perseveration	0.58 (63.74%)	1.58 (59.40%)	1.75 (35.64%)	2.50 (33.38%)
Metathesis	0.00 (0.00%)	0.08 (3.01%)	1.08 (22.00%)	0.16 (2.14%)
Inexplicable	0.33 (36.26%)	1.00 (37.59%)	2.08 (42.36%)	4.83 (64.48%)

Statistical analysis revealed a significant main effect for language group [$F(1, 44) = 12.57$, $p < 0.001$], i.e., the bilinguals made more "other" errors than did the monolinguals. Due to the small numbers in each of the "other" categories (perseveration, metathesis, and inexplicable), statistical analysis was not conducted to compare the occurrence of errors in these categories. However, at least one conclusion can safely be made: The bilinguals in the vocoded condition had more inexplicable errors (mean = 4.83) than did the bilinguals in the natural condition or the monolinguals in either the natural or vocoded condition.

2.5.6 Position-in-Sentence Errors

Analysis of these errors involved tabulating the number of erroneously rendered words occurring in each one of the 7 inter-sentential positions (see Figure 3). There was considerable



87633-3

Figure 3. Average number of position-in-sentence errors.

similarity in the overall error pattern among the monolinguals in the vocoded condition and the bilinguals in the natural and vocoded conditions. That is, there were almost no errors associated with words in position 1 (determiners), and there was a relatively large number of errors associated with words in positions 4 (verbs) and 6 (adjectives).

Moreover, for all subjects, there were more errors associated with determiners located in position 5 (near the end of the sentence) than with determiners in position 1 (in sentence-initial position). This finding cannot be accounted for solely on the basis of the fact that, in the stimulus set, there were 57 determiners in position 5 and 41 in position 1 — a ratio of 1.39 to 1.0. For subjects had, on the average, *4 to 10* times as many errors on determiners in position 5 as in position 1 — a value which greatly exceeds that of the 1.39-to-1.0 ratio. Likewise, there were far more errors associated with adjectives in position 6 than in position 2.

Additionally, the bilinguals in the vocoded condition performed similarly to the bilinguals in the natural condition with respect to words in positions 1, 2, and 3, but their error rates were higher on words in positions 4, 5, 6, and 7.

2.5.7 Task-Difficulty and Language-Proficiency Ratings

As the ratings in Table VI reveal, there was not a large difference in the four groups' task-difficulty ratings, although the bilinguals tended to rate the semantically anomalous test as being somewhat more difficult than the monolinguals did.

A Pearson product moment-correlation test was conducted to determine whether or not subjects' task-difficulty ratings correlated with their overall number of errors. Results indicated that the correlation was nonsignificant for the monolinguals in the natural condition. However, for the monolinguals in the vocoded condition, it was significant ($r = 0.53$, $p < 0.05$). It was also significant for the bilinguals in the natural condition ($r = 0.57$, $p < 0.05$) and in the vocoded condition ($r = 0.68$, $p < 0.01$). Thus, for three of the four test groups, there was a statistically significant positive correlation between task-difficulty rating and total number of errors. That is, subjects who had few errors tended to rate the test as easy; those who had many errors tended to rate it as difficult.

TABLE VI

Task Difficulty Ratings

	Monolinguals		Bilinguals	
	Natural	Vocoded	Natural	Vocoded
Mean	5.25	4.50	6.08	7.92
S.D.	1.66	2.47	2.81	1.88
Range	2-8	1-8	2-10	4-10

Tabulation of the bilinguals' English proficiency self-evaluation scores revealed that their average score was 14.48, with a range of 8.5 to 19.5. (Recall that the maximum score obtainable was 23.) Again, a Pearson product moment test was conducted to determine whether or not there was a significant correlation between the bilinguals' overall number of errors on the anomalous-sentences test and their English proficiency self-evaluation scores. Results revealed a significant negative correlation for the bilinguals in the natural condition ($r = -0.69$, $p < 0.01$) and in the vocoded condition ($r = -0.56$, $p < 0.05$). Thus, subjects who had few errors tended to rate themselves as more proficient in English; subjects who had many errors tended to rate themselves as less proficient.

3. DISCUSSION

The present study was designed to address two sets of questions posed in the Introduction — one set pertaining to a comparative analysis of monolingual and bilingual performance in response to natural and vocoded semantically anomalous sentences, and the other set pertaining to the relationship between selected subjective variables and test performance. In the following discussion, issues relevant to these questions will be addressed and suggestions for future research will be made.

3.1 COMPARISONS OF MONOLINGUAL AND BILINGUAL PERFORMANCE

3.1.1 Error Patterns

In response to the question, "In the present experiment, did the German-English bilinguals perform worse than the English monolinguals did?" an answer of yes must be given. The fact that the bilinguals did not, in general, perform as accurately as the monolinguals did is not especially surprising given that none of the bilinguals was a native speaker of English. What is somewhat surprising, however, is the fact that, in response to *natural* anomalous sentences, the bilinguals had an average of over 11 times as many errors as the monolinguals — in spite of the fact that all of the bilinguals were fluent speakers of English who were, at the time of testing, residing and working or studying in an English-speaking country.

Clearly, the demands of the task upon the bilinguals were considerable, suggesting that they usually rely heavily upon semantic, pragmatic, and contextual clues when they perceive English sentences. When such clues are absent, they experience a serious decrement in performance, even if the sentences are acoustically intact.

Also of interest is the fact that the vocoded sentences induced, relatively speaking, only slightly more errors for the bilinguals than the natural sentences did, while the vocoded sentences induced a considerable increase in the number of errors for the monolinguals. That is, the bilinguals made only 1.25 times as many errors on the vocoded as on the natural sentences, while the monolinguals made over 5 times as many. Apparently, responding to anomalous sentences was, of itself, so difficult for the bilinguals that distortion of the signal through vocoding yielded only a slight relative decrement in performance. Thus, the bilinguals' natural-sentence/vocoded-sentence error ratio was small not because they made so few errors on the vocoded sentences, but because they made so *many* on the natural sentences.

Of relevance to the notion that the semantically anomalous sentences were, of themselves, difficult for the bilinguals to process and/or recall is the relatively high number of "other" errors made by the bilinguals, both in the natural and vocoded conditions. The largest number of "other" errors were inexplicable (e.g., "chore" for "soap" and "vibrid" for "zebra") pointing, perhaps, to guessing.

To the claim that the bilinguals performed worse than did the monolinguals, this cautionary note must be added: In reporting only the overall number of errors and error ratios, we run the risk of concluding that the bilinguals performed *much* worse than the monolinguals did in both conditions. This may not be entirely true, as the percentage of correctly rendered words suggests.

This percentage, based upon the ratio of correctly rendered words to the total number of words in the stimulus set, indicates that the bilinguals' performance was not a great deal worse than the monolinguals'. (The bilinguals' percent correct was about 14 points less than the monolinguals' in the natural condition, and about 11 points less in the vocoded condition.) What must be recognized is that, even in response to vocoding, the bilinguals correctly rendered an average of 81.38% of the words in the stimulus set. Whether an average of 81% correct is deemed acceptable remains to be determined.

Note then that two approaches to scoring — the first consisting of overall number of errors and the second consisting of percent of correctly rendered words — can lead to two different interpretations. Use of the first scoring method indicates that the bilinguals did very badly on the semantically anomalous sentences test; use of the second scoring method suggests that they did fairly well. Clearly, scoring procedures should be examined carefully in tests such as this, and should be interpreted in light of the objectives and structure of the communication system which the intelligibility test is designed to assess. (For example, in the listening task, is it essential that nearly *every* phoneme of *every* word be perceived? Or is it sufficient that only *most* of the words be perceived? How serious are the consequences of misperception?)

A final and important observation regarding the pattern of errors is this: In the present experiment, on the average, from 32 to 45% of all errors were *not* phonemic. This clearly reveals that all linguistic components may be adversely affected in a perceptual experiment in which the stimuli are difficult to process. This should be kept in mind by researchers who rely solely upon tests designed to assess phoneme discrimination. While such tests may be adequate diagnostics, they cannot provide a complete picture of the extent to which the linguistic system *in toto* is affected by an acoustically degraded signal or a linguistically "vexing" stimulus.

3.1.2 Evidence of Linguistic Transfer

It was not one of the objectives of this experiment to examine the bilinguals' responses for evidence of linguistic transfer. However, there were so many instances of it that two observations can be made.

First, there were numerous instances of linguistic transfer (from German to English) in phonemic substitutions. For example, word-final obstruents are devoiced in German, resulting in, for example, the word "hand" being pronounced as [hant]. There were, accordingly, examples of word-final devoicing in the bilinguals' responses (but virtually none in the monolinguals'), yielding such productions as "seat" for "seed" and "cart" for "card." Another frequent source of errors was the interdental fricative /θ/, which is nonoccurrent in German. It was often converted to /s/, resulting in such productions as "sawed" for "thawed" and "sighs" for "thighs." While the

monolinguals also made such phonemic errors, especially in the vocoded condition, the bilinguals did so more frequently. Likewise, errors in the bilinguals' rendering of vowels may be attributable, at least in part, to differences in the vowel systems of German and English. Such productions as "Rax" for "Rex" and "mass" for "mess" were evident.

Second, the bilinguals often incorrectly parsed the English stimulus sentences. This often led to rather unusual constructions. Following are several examples of these:

Target	Response
"the bossy vapor"	"the boss evabor"
"High Mick thanked a zealous chin."	"Hi Mick! Thank a jealous chin!"
"a paper nature"	"a patronager"
"Modern Leslie"	"Maude and Leslie"

Differences in the segmental and prosodic cues for word-boundaries in German and English were most likely the cause for such misparsings. What is especially significant is that there were virtually no instances of misparsing among the English monolinguals.

3.1.3 Processing Strategies

Over the past 10 to 15 years, there has been considerable debate regarding the strategies utilized in speech processing. Much of this debate has centered upon the contrast between data-driven or "bottom-up" strategies and theory — or knowledge — driven or "top-down" strategies [30]. Due to the detail with which responses were analyzed in the present experiment, its findings can provide some insight into the strategies used by the monolinguals and bilinguals in transcribing the anomalous sentences.

In their 1978 article [31], Marslen-Wilson and Welsh (p. 30) make the following distinction between top-down and bottom-up processing.

> A top-down, or knowledge-driven system, . . . uses higher-level constraints on possible interpretations to "drive" the processing of the input data, in the sense that the system processes the data selectively, choosing interpretations that are consistent with these constraints. In a bottom-up or data-driven processor, it is the properties of the data themselves that are the primary determinants of the higher-level representation that the system proposes to account for the data.

Although numerous experiments have supported the notion that speech processing involves an interaction between both types of strategies, it must be recognized that certain types of stimuli, task demands, and listeners may determine which strategy is primarily utilized. For example, in perceiving vocoded semantically anomalous sentences in a potentially stressful testing situation, listeners may employ a strategy which is more data-driven than it would be if they

were processing natural meaningful sentences presented in the course of normal conversation. Likewise, it has been proposed that nonfluent second-language speakers utilize data-driven strategies to a greater extent than do highly fluent or native-language speakers [32].

Consistent with these premises is the notion that certain types of errors may be interpreted as evidence of top-down or bottom-up processing strategies. For example, Cziko [32] conducted a test involving the oral reading of French passages by intermediate and advanced level French students and by native speakers of French. He proposed that errors based upon the graphemic (physical) properties of the stimuli — such as the substitution of "many" for "money" — pointed to a reliance upon a data-driven strategy, while errors based upon the syntactic or semantic structure of the stimuli — such as the substitution of "dimes" for "money" — pointed to a reliance upon a knowledge-driven strategy. Further, Cziko found evidence that the native speakers of French made more errors based upon the syntactic and semantic structure of the stimuli than did the advanced or intermediate speakers, whose errors were more often based upon the phonological or graphemic properties of the stimuli.

If Cziko's observations can be extended to auditory stimuli, then we might predict that the German-English bilinguals in the present study would have relatively more phonemic errors than the monolinguals, revealing data-driven processing. We would likewise predict that relatively more phonemic errors would emerge in response to vocoded than to natural sentences, that is, in response to stimuli which are more difficult to process.

Analysis of the data reveals that the monolinguals and bilinguals may have utilized similar processing strategies, but that the bilinguals simply found the task, overall, more difficult. Specifically, for all four groups of subjects, and in both listening conditions, phonemic errors predominated, possibly pointing to the fact that the anomalous sentences encouraged all subjects to utilize a data-driven processing strategy. On the other hand, while the monolinguals and bilinguals had approximately the same percentage of phonemic errors in response to natural speech, the bilinguals had nearly 10% *fewer* phonemic errors than the monolinguals did in response to vocoded speech. While this could be interpreted as evidence of a more top-down processing strategy among these subjects, it may actually reflect the basic difficulty experienced by the bilinguals in processing the vocoded sentences, for they also had a high percentage of lexico-semantic errors, and over 97% of these lexico-semantic errors were omissions. (Because the percentages reflect relative distributions, as the percentage of lexico-semantic errors rises, the percentage of phonemic errors falls.) Hence, while the error analysis does not provide compelling evidence that the bilinguals utilized a processing strategy which differed from that of the monolinguals', it is evident that the bilinguals had considerable difficulty in processing the sentences; the high percentage of lexico-semantic omissions reveals that they did not understand and/or could not recall all of the stimulus words.

This suggests that lexical access and/or recall suffered as a function of task difficulty. This notion is supported by studies conducted by Rabbit [33] and Luce et al. [6]. These researchers demonstrated that subjects experienced decrements in retention of linguistic stimuli when those stimuli were acoustically degraded. If omissions may be attributed to problems in recall, then it is

clear that the bilinguals in the vocoded condition had the greatest recall difficulties, for they had the highest number of lexical-semantic (content-word) omissions. The capacity demands on their processing systems may have been so great that they simply had difficulty retaining and reproducing every word presented.

The notion that subjects experienced difficulty in recall receives additional support from the "position-in-sentence" error analysis. Specifically, determiners in position 1 exhibited fewer errors than determiners in position 5, and adjectives in position 2 exhibited fewer errors than adjectives in position 6.

An additional point, concerning the patterns of response, is that the number of substitution errors exceeded that of insertion and omission errors for subjects in all four groups. Thus, when subjects made errors, they tended to retain the segmental structure of the stimulus sentences, even if the specific phonemes they reproduced were erroneous.

Finally, it is important to recognize that errors in the present task could reflect any combination of the following factors: (1) degradation in the acoustic quality of the stimuli; (2) misperception of acoustically accurate stimuli; and (3) inability to recall the stimuli correctly. Evidence from the present study suggests that all three of these factors influenced subjects' performance. Vocoded speech yielded more errors than natural speech did; the bilinguals made many errors even in response to natural speech; and determiners and adjectives located in (or near) sentence-initial position were more accurately reproduced than determiners and adjectives near sentence-final position.

3.1.4 Correlations Between Subjective Evaluations and Test Scores

In addition to addressing the above issues, the present study was designed to explore the relationship between subjects' self-evaluations and their test scores. In operational terms, the self-evaluations consisted of subjects' task-difficulty ratings and, for the bilinguals, English-proficiency ratings. The test scores consisted of overall number of errors on the semantically anomalous sentences test.

Correlational analysis revealed significant correlations between task-difficulty ratings and test scores for three of the four test groups, with the only exception being the English monolinguals in the natural-speech condition. It may have been that these subjects simply found the natural-speech stimuli so easy to respond to that their ratings were not a valid indicator of their test performance.

In light of this finding, however, it does appear that subjective ratings are, at least for a test such as this, only moderately valid indicators of task difficulty. Actual test performance is clearly a more sensitive measure of difficulty. Note that the monolinguals in the vocoded condition actually rated the test as being somewhat *easier* — with a mean rating of 4.50 — than did the monolinguals in the natural condition — with a mean rating of 5.25. And yet the monolinguals in the vocoded group had an average of over 5 times as many errors as the monolinguals in the natural group.

Correlational analysis of the bilinguals' subjective English proficiency ratings and their test scores was also carried out. While the correlations were not very strong (-0.69 for those in the natural condition and -0.56 for those in the vocoded condition), they were statistically significant. Thus, such ratings may be useful indicators of language proficiency.

The correlation could have been strengthened, perhaps, if the proficiency rating incorporated subjective evaluations with objective variables (e.g., age at the onset of exposure to English, number of years spent studying English, etc). Nonetheless, the fact that there were significant correlations between subjects' test scores and their subjective evaluations of English proficiency suggests that, in the absence of other background data, such evaluations may be sufficient indicators of L2 ability, and hence can be used for, e.g., placing subjects in different proficiency groups for testing purposes.

3.2 SUGGESTIONS FOR FURTHER RESEARCH

The two major findings of the present experiment were that (1) non-native speakers of English did not respond as accurately to English semantically anomalous sentences as English monolinguals did, either when these sentences were natural or vocoded; and (2) both native and non-native speakers of English responded less accurately to vocoded sentences than to natural sentences. These findings carry some important implications for future research.

Several obvious questions, for example, are these: Is the performance of this group of non-native speakers characteristic of others? That is, would native speakers of languages other than German perform similarly? Further, if the bilinguals had all acquired English in early childhood, rather than in adolescence or early adulthood, would they have performed better? These questions could be addressed in studies in which specific language groups and subject types are examined. This seems quite important, given that previous research has demonstrated that bilinguals' native language and language proficiency influences their perception of L2 stimuli.

Related to the above questions is the possibility that, if specific types of error patterns emerge among different language groups (as they no doubt would), the acoustic properties of the synthetic stimuli could be selectively enhanced to improve specific features for the particular listeners to whom the speech signal was being directed.

In addition, it must be asked whether or not the results of anomalous-sentence tests are consistent with data obtained from other types of tests. The results of Greene et al. [21] and Mack and Gold [9] suggest that they are. If this is the case, then tests which are easier to score (such as the DRT) could be used in place of anomalous-sentence tests. But what such tests usually do not, or indeed cannot do, is demonstrate the extent to which *nonphonemic* errors occur when perception is impaired. Implicit in the DRT, for example, is the assumption that intelligibility is an acoustic- and phonetic-based bottom-up process, and that high scores on this test mean that acceptable levels of perceptual performance have been obtained. Yet, this can only be concluded by comparative analyses of DRT and other test results, especially those involving sentence processing. (Recall that, in the present study, about 32 to 45% of the errors were not

phonemic.) Hence, in order to assess the intelligibility of any speech system, it seems important to evaluate it using two or more types of tests, at least until the relationship between various test results and "real-life" perceptual performance is more completely understood.

This raises a final, and extremely important issue. One criticism levied against many experiments in speech perception and intelligibility is that they are not "ecologically valid," i.e., that their stimuli, task demands and/or modes of response are unlike those used in the actual conditions which the experiments are designed to assess. Yet the reason for administering tests which seem to lack external or ecological validity are obvious: There is a degree of control over subjects and stimuli which is often impossible to obtain in more naturalistic contexts. This control, of course, permits the experimenter to infer causality through the manipulation of experimental variables. Yet, once such experiments have been conducted, it is important to attempt to relate their results to actual nonexperimental contexts.

A reasonable approach then would be to test non-native speakers (listeners) using stimuli and listening conditions with which they are normally confronted. It may be, for example, that a test such as the one used in the present experiment is unfairly biased against non-native speakers. On the other hand, such a test may, by placing excessive demands upon subjects' perceptual mechanisms and recall abilities, be an especially sensitive indicator of linguistic competence. What is needed are comparative assessments of non-native subjects' responses in highly controlled and naturalistic research settings. For example, it would be appropriate to test the perceptual performance of the recipients of vocoded speech using the types of messages, vocoders, and listening conditions with which these recipients are commonly confronted. In this way, we could better determine whether or not the performance levels demonstrated in the present study are acceptable, or whether they are indicative of potentially serious problems in the intelligibility of synthetic speech perceived by non-native listeners.

4. CONCLUSION

The present study was designed to evaluate the performance of 24 English monolinguals and 24 German-dominant German-English bilinguals in response to 57 natural and computer-generated (vocoded) semantically anomalous sentences. The performance data were quantified in a number of ways. These included: (1) mean number of overall errors; (2) percentage of correctly rendered words; (3) mean number and percentage of linguistic, transpositional, and "other" errors; and (4) mean number of errors associated with each within-sentence position. In addition, correlational analyses were conducted to determine the strength of the relationship between subjects' task-difficulty ratings and their test performance and between the bilinguals' English-proficiency evaluations and their test performance.

Results revealed that the bilinguals made far more errors than the monolinguals did in response to natural and vocoded speech, but that the ratio of vocoded-speech-to-natural-speech errors was considerably larger for the monolinguals than for the bilinguals.

For all groups of subjects, phonemic errors were the most predominant. Yet approximately 30-40% of all errors were not phonemic, indicating that other linguistic components were adversely affected as well. In addition, most errors were phonemic substitutions (e.g., "thighs" → "sighs"), rather than omissions or insertions. This demonstrates that subjects successfully maintained the essential structure of the stimulus sentences. However, the bilinguals in the vocoded condition exhibited numerous lexico-semantic omissions (i.e., deletion of content words), perhaps due to excessive demands on processing and/or recall.

Subjects in all test groups made, on the average, fewer errors in response to sentence-initial words than in response to sentence-final words. This finding is interpreted as reflecting potentially significant recall effects. Such effects are not apparent in conventional speech-perception tests (such as the DRT) which utilize single-word stimuli and hence place minimal demands on memory.

There was no clear evidence that the bilinguals and monolinguals utilized different processing strategies. Phonemic errors (and, more specifically, phonemic substitutions) predominated for all groups, suggesting that all found it difficult to apply a top-down strategy in processing the anomalous sentences.

Significant positive correlations were found between task-difficulty ratings and number of overall errors for three of the four test groups. Significant negative correlations were found between the bilinguals' English-proficiency ratings and number of overall errors. These findings suggest that subjective ratings *may* be useful diagnostic tools in language research, although they should probably be used in conjunction with objective data.

Topics considered important for future research include the following: (1) determining to what extent the present findings can be generalized to subjects from other language groups and to bilingual subjects at different levels of English ability; (2) considering the possibility of enhancing certain acoustic/phonetic properties of the speech signal to improve intelligibility for specific language groups; and (3) comparing performance on highly-controlled experimentally-based intelligibility tests with performance on more naturalistic and ecologically valid tests.

REFERENCES

1. W.D. Voiers, "Diagnostic Evaluation of Speech Intelligibility," in *Speech Intelligibility and Speaker Recognition*, ed. by M.E. Hawley (Dowden, Hutchinson and Ross, Stroudsburg, Pennsylvania, 1977), pp. 374-387.

2. D.B. Pisoni and S. Hunnicutt, "Perceptual Evaluation of MITalk: The MIT Unrestricted Text-to-Speech System," 1980 IEEE International Conference Record on Acoustics, Speech and Signal Processing, (1980) pp. 572-575.

3. D.B. Pisoni, "Speeded Classification of Natural and Synthetic Speech in a Lexical Decision Task," J. Acoust. Soc. Am., **70**, S98 (1981).

4. D.B. Pisoni and E. Koen, "Some Comparisons of Intelligibility of Synthetic and Natural Speech at Different Speech-to-Noise Ratios," J. Acoust. Soc. Am., **71**, S94 (1982).

5. B. Gold and J. Tierney, "Vocoder Analysis Based on Properties of the Human Auditory System," Technical Report 670, Lincoln Laboratory, MIT (22 December 1983) DTIC AD-A138660/6.

6. P.A. Luce, J.C. Feustel, and D.B. Pisoni, "Capacity Demands in Short-Term Memory for Synthetic and Natural Speech," Human Factors, **25**, 17-32 (1983).

7. D.B. Pisoni, L.M. Manous, and M.J. Dedina, "Comprehension of Natural and Synthetic Speech: II. Effects of Predictability on the Verification of Sentences Controlled for Intelligibility," Speech Research Laboratory, Progress Report 12, Indiana University (1986).

8. J.J. Jenkins and L.D. Franklin, "Recall of Passages of Synthetic Speech," presented at Psychonomics Society Meeting (1981).

9. M. Mack and B. Gold, "The Intelligibility of Non-Vocoded and Vocoded Semantically Anomalous Sentences," Technical Report 703, Lincoln Laboratory, MIT (26 July 1985), DTIC AD-A160401.

10. K.N. Stevens, A.M. Liberman, M. Studdert-Kennedy, and S.E.G. Öhman, "Cross-Language Study of Vowel Perception," Language and Speech, **12**, 1-23 (1969).

11. A. Caramazza, G.H. Yeni-Komshian, E.B. Zurif, and E. Carbone, "The Acquisition of a New Phonological Contrast: The Case of Stop Consonants in French-English Bilinguals," J. Acoust. Soc. Am., **54**, 421-428 (1973).

12. M. Schouten, *Native-Language Interference in the Perception of Second-Language Vowels* (Drukkerijelinkwijk B.V., Utrecht, 1975).

13. S. Garnes, "Some Effects of Bilingualism on Perception," Papers in Psycholinguistics and Sociolinguistics, Ohio State University (1977).

14. L. Williams, "The Modification of Speech Perception and Production in Second Language Learning," Percept. & Psychophys. **26**, 95-104 (1979).

15. K. MacKain, C. Best, and W. Strange, "Categorical Perception of English /r/ and /l/ by Japanese Bilinguals," Applied Psycholing, **2**, 369-390 (1981).

16. M. Mack, "Voicing-Dependent Vowel Duration in English and French: Monolingual and Bilingual Production," J. Acoust. Soc. Am., **71**, 173-178 (1982).

17. M. Mack, "Early Bilinguals: How Monolingual-Like Are They?," in *Early Bilingualism and Child Development*, ed. by M. Paradis and Y. Lebrun (Swets & Zeitlinger, Lisse, 1984), pp. 161-173.

18. B. Voss, *Slips of the Ear: Investigations into the Speech Perception Behaviour of German Speakers of English* (Gunter Narr Verlag, Tubingen, 1984).

19. J.E. Flege and J. Hillenbrand, "Differential Use of Temporal Cues to the /s/-/z/ Contrast by Native and Non-Native Speakers of English," J. Acoust. Soc. Am., **79**, 508-517 (1986).

20. A.K. Nábělek and A.M. Donahue, "Perception of Consonants in Reverberation by Native and Non-Native Listeners," J. Acoust. Soc. Am., **75**, 632-634 (1984).

21. B.G. Greene, D.B. Pisoni, and H.L. Gradman, "Perception of Synthetic Speech by Non-Native Speakers of English," Speech Research Laboratory, Progress Report 11, Indiana University (1985).

22. M. Mack, "Psycholinguistic Consequences of Early Bilingualism: A Comparative Study of the Performance of English Monolinguals and French-English Bilinguals in Phonetics, Syntax, and Semantics Experiments," Ph.D. Thesis, Brown University (1983).

23. P. Lieberman, "Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech," Language and Speech, **6**, 172-187 (1963).

24. I. Pollack and J.M. Pickett, "Intelligibility of Excerpts from Conversational Speech," Language and Speech, **6**, 165-173 (1963).

25. R. Reddy, "Speech Recognition by Machine: A Review," Proc. of the IEEE, **64**, 501-531 (1976).

26. D.H. Klatt, "Review of the ARPA Speech Understanding Project," J. Acoust. Soc. Am., **62**, 1345-1366 (1977).

27. B.J. Winer, Statistical Principles in Experimental Design, 2nd ed. (McGraw-Hill, New York, 1971).

28. G.A. Ferguson, *Statistical Analysis in Psychology and Education*, 5th ed. (McGraw-Hill, New York, 1981).

29. J.W. Tukey, "Comparing Individual Means in the Analysis of Variance," Biometrics, **5**, 99-114 (1949).

30. J. Ohala, "Phonological Evidence for Top-Down Processing in Speech Perception," *Invariance and Variability in Speech Processes*, ed. by J.S. Perkell and D.H. Klatt (Lawrence Erlbaum, Hillsdale, New Jersey, 1986) pp. 386-401.

31. W.D. Marslen-Wilson and A. Welsh, "Processing Interactions and Lexical Access During Word Recognition in Continuous Speech," Cog. Psych. **10**, 29-63 (1978).

32. G.A. Cziko, "Language Competence and Reading Strategies: A Comparison of First- and Second-Language Oral Reading Strategies," Language Learning, **30**, 101-114 (1980).

33. P. Rabbitt, "Recognition Memory for Words Correctly Heard in Noise," Psychon. Science, **6**, 383-384 (1966).

APPENDIX I

Language-Background Questionaire, Part 1

Name:

Age:

You are currently a citizen of what country?

Your native language:

Age at onset of acquisition of English:

What other languages do you speak besides English and German?

Highest level of school completed:

School currently attending:

Mother's native language:

Father's native language:

Language spoken at home by mother:

Language spoken at home by father:

Your preferred language:

What other members of your family, if any, can speak English?

What language do you count in?

What language do you say the alphabet in?

When you talk silently to yourself, what language do you use?

Briefly describe how you acquired English:

Language-Background Questionnaire, Part 2

1. Are there grammatical structures in English which you try to avoid using?
2. Do you sometimes find yourself in the middle of an English sentence you cannot finish because of limitations in your grammar or vocabulary?
3. Do you find it difficult to follow and contribute to a conversation among native speakers of English who try to include you in their talk?
4. Are you afraid that you will misunderstand information given to you in English over the phone?
5. Do native speakers of English ever correct your pronunciation?
6. Do native speakers of English ever correct your sentence structure?
7. Could you serve as an interpreter for a government official on all diplomatic and social functions?
8. Do you think that you practically never mak. mistakes in English?
9. Do native speakers of English normally speak to you in their own language?
10. Do you believe you speak with any trace of a German accent in English?
11. In long conversations, are you taken for a native speaker of English?

    never _____
    sometimes _____
    always _____

12. Can you talk about anything in English easily?
13. Rate your overall proficiency in English on a scale of 1-10.
    (1 represents the score of a low beginner and 10 represents the score of a native speaker)
14. What is your TOEFL score?

APPENDIX II

Semantically Anomalous Sentences

1. A painted shoulder thawed the misty sill.
2. The bitter seed vexes a valid dinner.
3. The tacky runner judged a short fact.
4. Dingy Doug chips the poor jewel.
5. A golden corner varies the thoughtful keeper.
6. A cotton zebra thickened the chief tickle.
7. The simple rocket picks a new female.
8. A zesty joke gets the nice feather.
9. The shiny shore gives a heavy father.
10. Checkered Sharon gained the chilly hope.
11. Recent Gary sets a messy shower.
12. Fake Chuck finished the hopeful golfer.
13. The vague job savors a jolly garden.
14. A thin jailer checked a meager soap.
15. Moody Tim holds the sane zero.
16. A newer deed shines a safe sinner.
17. A luscious devil helps the good raid.
18. The jealous duster lifted a gaudy cap.
19. The helpful knitter makes a gabby lip.
20. A paper nature seeks the cool master.
21. The bossy vapor shakes a careful victor.
22. Top Jane zapped the tense tot.
23. A dark nail zones the round reason.
24. The kind ladder shoots a dim bed.
25. The gilded nest zipped the dusty tank.
26. The zingy thing liked a late toddler.
27. The soft bargain mixes a thick needle.
28. A shoddy lobby mopped the dense hip.

29. Modern Leslie healed a cheap hat.
30. The charming deck robbed the hot jelly.
31. A jaunty fork raised a vacant cow.
32. The funny heaven reads the shallow pepper.
33. Ready Holly doubts the shabby van.
34. Novel Cathy dipped the loud hopper.
35. A vain foam denies a zippy lime.
36. The third pattern teases a zany tailor.
37. High Mick thanked a zealous chin.
38. Healthy Ned tears the solid rat.
39. Lean Rex takes the pale chowder.
40. A lewd pill leads a pink zing.
41. The bizarre pot needed the best zombie.
42. A partial baker knocked the boring shell.
43. Tipsy Peter keeps the better chopper.
44. The damp vase catches a tiny zeal.
45. A kingly thinker bites a nasty lock.
46. A gorgeous villain chopped the rotten thimble.
47. The southern gift beats the tall thighs.
48. Sure Susan bought a famous thirst.
49. A jagged sailor paid a ripe card.
50. A cheerful thistle pours the fat bean.
51. The zinc mitt carries a lazy basket.
52. A feisty chain fights the fertile money.
53. Vast Bob jabbed a junior pack.
54. The thirsty vine finds a giant shop.
55. The moral gold vacates a costly gate.
56. A normal cheater joined the thorough mess.
57. Rapid Zach nabs a vulgar mirror.

ACKNOWLEDGMENTS

We wish to thank Clifford Weinstein for his cooperation and encouragement on this project. Much gratitude is also extended to Ben Gold whose work in vocoding the sentences was invaluable.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report 792	5. MONITORING ORGANIZATION REPORT NUMBER(S) ESD-TR-87-081
6a. NAME OF PERFORMING ORGANIZATION Lincoln Laboratory, MIT 6b. OFFICE SYMBOL *(If applicable)*	7a. NAME OF MONITORING ORGANIZATION Electronic Systems Division
6c. ADDRESS *(City, State, and Zip Code)* P.O. Box 73 Lexington, MA 02173-0073	7b. ADDRESS *(City, State, and Zip Code)* Hanscom AFB, MA 01731
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Rome Air Development Center 8b. OFFICE SYMBOL *(If applicable)* RADC/EEV	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F19628-85-C-0002
8c. ADDRESS *(City, State, and Zip Code)* Hanscom Air Force Base Massachusetts 01731	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 33401F, 64754F; PROJECT NO; TASK NO; WORK UNIT ACCESSION NO

11. TITLE *(Include Security Classification)*
The Intelligibility of Natural and Vocoded Semantically Anomalous Sentences: A Comparative Analysis of English Monolinguals and German-English Bilinguals

12. PERSONAL AUTHOR(S)
Molly Mack (University of Illinois) and Joseph Tierney

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT *(Year, Month, Day)*	15. PAGE COUNT
Technical Report	FROM _____ TO _____	10 December 1987	46

16. SUPPLEMENTARY NOTATION
None

17. COSATI CODES			18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
FIELD	GROUP	SUB-GROUP	bilingual, vocoder, speech perception, German, English, computer-generated, speech synthesis, intelligibility, second-language acquisition, speech processing

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The present study was undertaken in order to analyze the performance of 24 German-dominant German-English bilinguals and 24 English monolinguals on tests of semantically anomalous natural and computer-generated (vocoded) sentences. The primary objectives of the study were these: to determine whether the overall performance of the bilinguals was significantly worse than the monolinguals in response to natural and/or vocoded speech; to categorize the specific types of errors made by the two groups of subjects; and to analyze the patterns of errors made by the two groups in an attempt to evaluate their sentence-processing strategies. Secondary objectives of the study were these: to assess the relationship between the groups' subjective ratings of task difficulty and their test performance; and to assess the relationship between the bilinguals' subjective evaluation of their English proficiency and their test performance.

Results revealed that, overall, the bilinguals made more errors than the monolinguals did in response to both natural and vocoded speech. Further, both groups had a preponderance of phonemic rather than morpho-syntactic or lexico-semantic errors, and most errors were phonemic substitutions, rather than omissions or insertions. Results of the analysis suggested that the bilinguals and monolinguals were using similar processing strategies. The bilinguals' pattern of errors indicated that they found both the natural and vocoded sentence tasks considerably more difficult than the monolinguals did, and the number of overall errors in the vocoded sentence task indicated a potential problem for communication systems of this type. It was also found that subjective task-difficulty ratings and English-proficiency ratings were correlated with test scores.

Practical and theoretical implications of the results are considered, and suggestions for additional research are provided.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt. Col. Hugh L. Southall, USAF	22b. TELEPHONE *(Include Area Code)* (617) 863-5500, Ext. 2330 22c. OFFICE SYMBOL ESD/TML

END

DATE

FILMED

MARCH

1988

DTIC